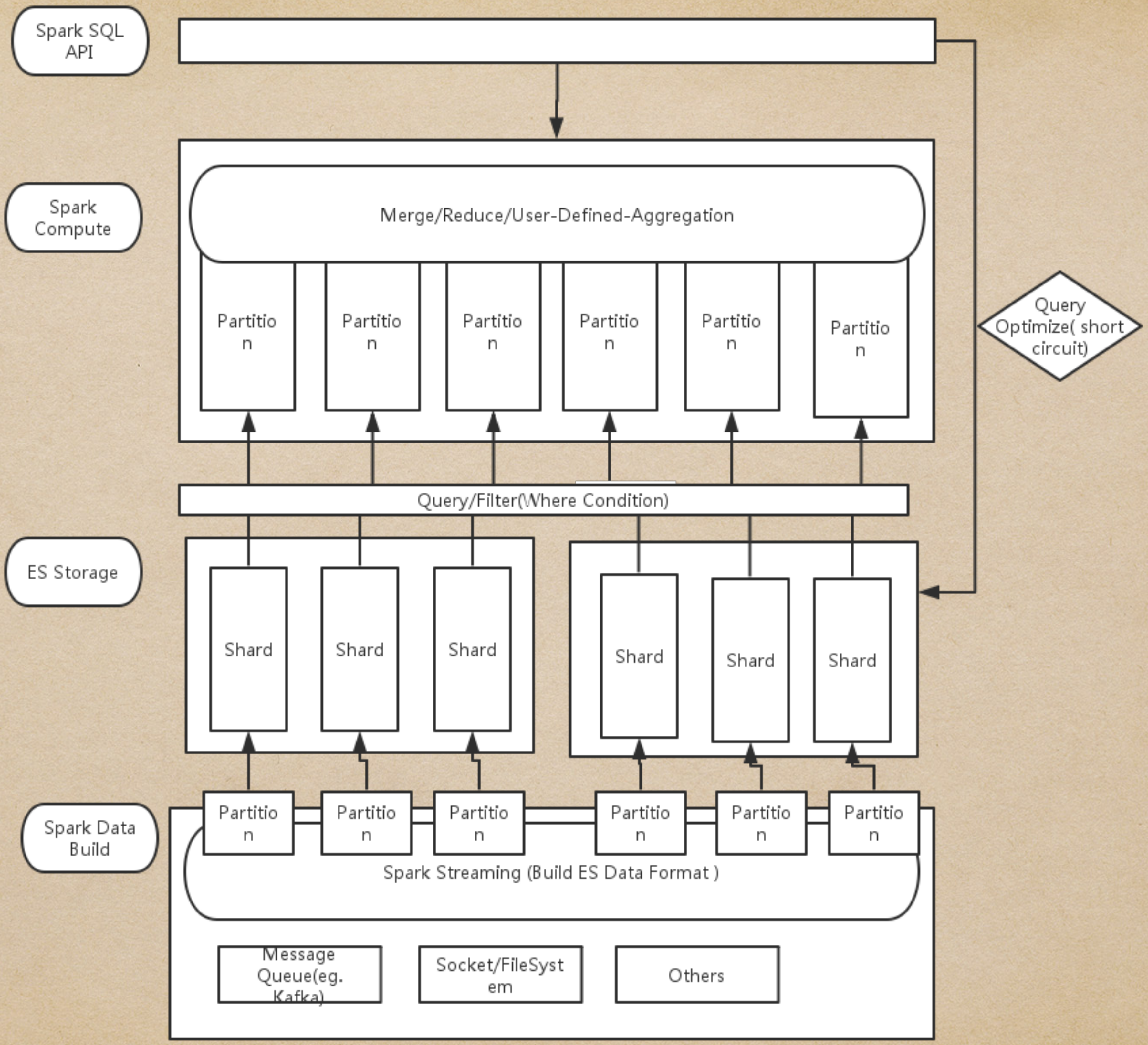
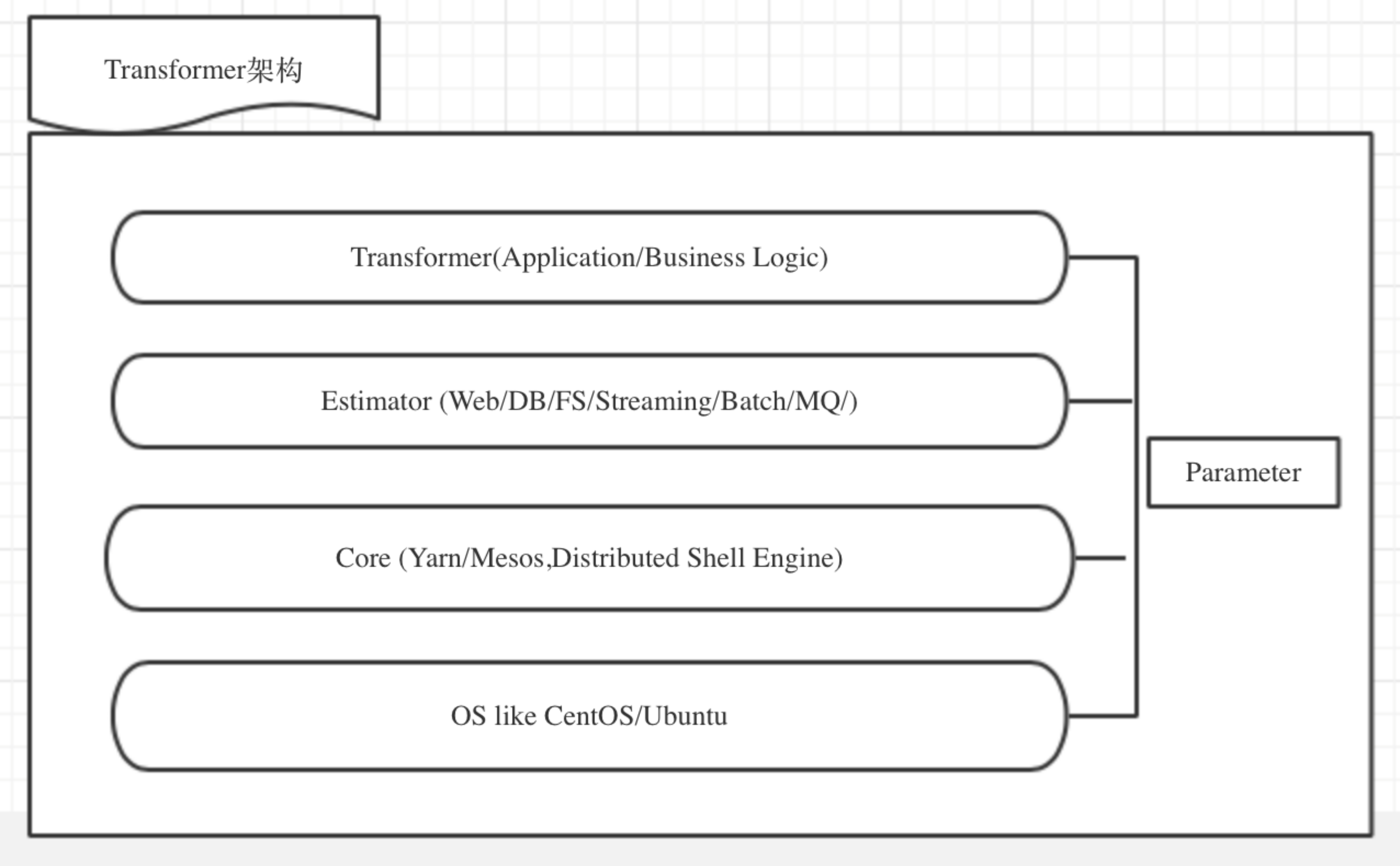
Integration between ElasticSearch and Spark

祝威廉

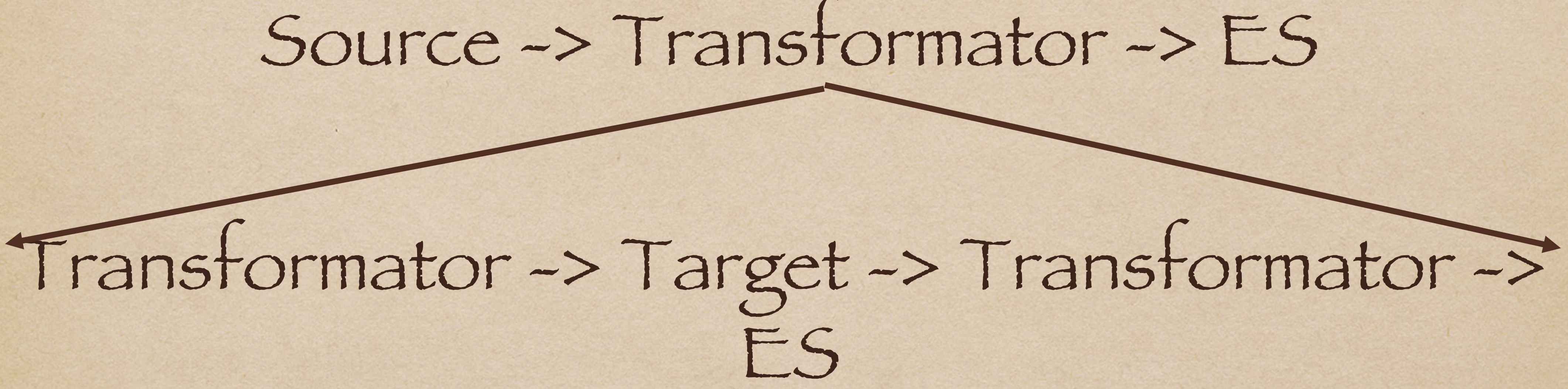

Spark SQL API
Spark Compute
Merge/Reduce/User-Defined-Aggregation
Partitio n
Partitio n
Partitio n
Partitio n
Partitio n
Partitio n
Query Optimize(short circuit)
Query/Filter(Where Condition)
ES Storage
Shard
Shard
Shard
Shard
Shard
Shard
Spark Data Build
Partitio n
Partitio n
Partitio n
Partitio n
Partitio n
Partitio n
Spark Streaming (Build ES Data Format)
Message Queue(eg. Kafka)
Socket/FileSyst em
Others


Transformer架构
Transformer(Application/Business Logic)
Estimator (Web/DB/FS/Streaming/Batch/MQ/)
Core (Yarn/Mesos,Distributed Shell Engine)
OS like CentOS/Ubuntu
Parameter

multi connected Transformers
construct data Pipeline

Tranformation Architecture

More detail about transformation Architecture
please visit:
http://www.jianshu.com/p/8a88a8bb4700



Source -> Transformator -> ES

Transformator -> Target -> Transformator -> ES

Step one: Pipeline of Pushing your data to ES

How to build pipeline of pushing your data to ES

- We choose Spark Streaming as Estimator
- Spark Streaming support many sources,eg. Kafka,Socket,HDFS/S3,Kinesis/Twitter
- ElasticSearch-hadoop project connect Spark Streaming and ES
- Spark Streaming have powerful transformation operator to do what you want on data

Write SQLs, Chain them，Submit，Done!

so we create a project called StreamingPro make it easy to transform and put data to ES

More detail about StreamingPro
please visit:
https://github.com/allwefantasy/streamingpro

How to query ES

- Query ES directly based on es-sql
- Query ES from Spark based on es-hadoop
- Combine them with optimize engine
- Shard to partition (optimize loading data from ES)

Optimize Engine

- Analysis of SQL to choose which action to take
- Load Aggregation from ES (Not Supported by ES-Hadoop,ES now is Compute Engine and Storage)
- Load Data from ES (ES is just storage)
- Direct Mode

Spark & ES

- Make Spark do more computation
- Make full use of ES Aggregation and Full Index Search
- Append only Segment

Improve ES bulk performance

- Optimize ES configurations(like bulk threads)
- Optimize es-hadoop configurations(like byteSize/ batchSize)
- enable AutoIdGeneration (2.1.0 having bug)
- Partition to Shard optimize (es-hadoop)
- ExternalAutoIdGeneration(es-hadoop/es)
- Flush manually/Increment Shard Num

Improve ES query performance

- Leave more memory to OS file cache (docValues/segments)
- More disks /More shards
- Move aggregation reduce to Spark

ES have Spark， fire now!

–祝威廉

Be thankful